MOLECULAR ECOLOGY
Molecular Ecology (2014) **23**, 5538–5551 doi: 10.1111/mec.12948

Synechococcus: 3 billion years of global dominance

PETR DVOŘÁK,* DALE A. CASAMATTA,† ALOISIE POULÍČKOVÁ,* PETR HAŠLER,* VLADAN ONDŘEJ* and REMO SANGES‡
*Department of Botany, Faculty of Science, Palacký University Olomouc, Šlechtitelů 11, CZ-78371 Olomouc, Czech Republic, †Department of Biology, University of North Florida, 1 UNF Drive, Jacksonville, FL 32224, USA, ‡Laboratory of Animal Physiology and Evolution, Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Napoli, Italy

Abstract

Cyanobacteria are among the most important primary producers on the Earth. However, the evolutionary forces driving cyanobacterial species diversity remain largely enigmatic due to both their distinction from macro-organisms and an undersampling of sequenced genomes. Thus, we present a new genome of a *Synechococcus*-like cyanobacterium from a novel evolutionary lineage. Further, we analyse all existing 16S rRNA sequences and genomes of *Synechococcus*-like cyanobacteria. Chronograms showed extremely polyphyletic relationships in *Synechococcus*, which has not been observed in any other cyanobacteria. Moreover, most *Synechococcus* lineages bifurcated after the Great Oxidation Event, including the most abundant marine picoplankton lineage. Quantification of horizontal gene transfer among 70 cyanobacterial genomes revealed significant differences among studied genomes. Horizontal gene transfer levels were not correlated with ecology, genome size or phenotype, but were correlated with the age of divergence. All findings were synthetized into a novel model of cyanobacterial evolution, characterized by serial convergence of the features, that is multicellularity and ecology.

Keywords: 16S rRNA, cyanobacteria, evolution, genome, horizontal gene transfer, speciation

Received 23 July 2014; revision received 23 September 2014; accepted 26 September 2014

Introduction

Synechococcus is widely considered one of the most abundant photo-oxygenic micro-organisms on Earth (Whitton & Potts 2000). Together with *Prochlorococcus*, it is responsible for ca. 25% of oceanic net primary production (Flombaum *et al.* 2013). *Synechococcus* is a cosmopolitan genus of cyanobacteria found in marine, freshwater, terrestrial and subaerial habitats. Further, its range extends from arctic to tropical waters, as an epiphyte, free living or in symbiotic relations with plants and animals (Honda *et al.* 1999; Robertson *et al.* 2001; Usher *et al.* 2004; Erwin & Thacker 2008; Dvořák *et al.* 2014). While ubiquitous in nearly all habitats, *Synechococcus* is most intensively studied as a common component of pelagic marine picoplankton (Haverkamp *et al.* 2009) and thermal habitats (Papke *et al.* 2003).

Synechococcus taxa are small (ca. 0.4–6 μm), bacilloid organisms, capable of forming pseudo-filaments, with little additional distinguishing morphology. It has long been considered as a coherent genus due to low morphological variability, which is also apparent in the ultrastructure of the cells (Komárek *et al.* 1999). This greatly complicates taxonomical differentiation, because early molecular techniques showed significant differences among *Synechococcus* strains, for example in G+C content (Waterbury & Rippka 1989), suggesting diverse evolutionary trajectories and thus separation to different genera. Some of these genera have already been described, for example *Cyanobium* and *Cyanobacterium* (Holt *et al.* 1994), but the evolutionary relationships within the *Synechococcus* group seem to be significantly more entangled. Recently, investigations employing genetic markers (e.g. the 16S rRNA and 16S-23S ITS gene regions) have noted at least five (Honda *et al.* 1999) and as many as seven (Robertson *et al.* 2001) clades, which do not seem to have a corresponding ecological or morphological signature. For example, hot spring taxa have been recovered in multiple lineages based on sequence data, and freshwater and marine taxa are likewise interspersed throughout the lineages.

Correspondence: Petr Dvořák, Fax: +420 585 634 824;
E-mail: p.dvorak@upol.cz

Moreover, new *Synechococcus*-like evolutionary lineages are still being discovered, for example *Neosynechococcus* (Dvořák *et al.* 2014).

One of the principal difficulties in reconstructing prokaryotic phylogenetic lineages is elucidating the evolutionary pressures on lineage divergence. As cyanobacteria are asexual, many of the evolutionary models employed by other taxonomic groups are inappropriate. However, genetic exchange between cyanobacterial lineages is possible and may be quite common. For example, both horizontal gene transfer (HGT) (Zhaxybayeva *et al.* 2006; David & Alm 2010; Polz *et al.* 2013) and homologous recombination (HR, Fraser *et al.* 2007) have been postulated as significant evolutionary forces in cyanobacteria. While some have proposed that bacteria with significant HR (when the HR exceeds mutation rates) are in fact sexual (Fraser *et al.* 2007), numerous competing models have advocated otherwise (see Polz *et al.* 2013 for review).

Due to their ubiquity and ecological significance, the majority of research relating to *Synechococcus* has thus far been focused on marine habitats. Recent research has uncovered a bounty of cryptic diversity, with speciation being evidenced by differential photopigment production not necessarily reflected in 16S rRNA diversity (Rocap *et al.* 2002). The genus as a whole is increasingly being investigated on a genomic scale, with 26 of 43 currently completely sequenced cyanobacterial genomes being from *Synechococcus* (GOLD database, http://genomesonline.org/cgi-bin/GOLD/index.cgi). However, there exists a dearth of information correlating genomic data with ecology or evolutionary history within these lineages, and nearly none from nonmarine environments. The work presented herein contributes to this knowledge gap by providing: (i) a de novo assembly of the genome of a novel *Synechococcus*-like strain from a peat bog, (ii) an assessment of phylogenies of available *Synechococcus* 16S rRNA sequences comparing them to trees generated from current available *Synechococcus* genomes, (iii) an estimate of the timescale of *Synechococcus*-like cyanobacteria evolution using molecular clocks and (iv) an examination of the potential evolutionary forces shaping the speciation events (e.g. the role and tempo of HGT).

Methods

Draft genome sequencing

Strain description and culture conditions are available in Dvořák *et al.* (2014). The strain *Neosynechococcus sphagnicola* CAUP A 1101 has been isolated from a peat-bog Klin (19°29′E, 49°25′N) in Protected Landscape Area Horná Orava, near Namestovo (Slovakia). It inhabits hyaline cells of *Sphagnum* sp., sheaths of cyanobacteria and other surfaces with biotic origin, and solitary in detritus. The strain was unialgal with bacterial contamination and maintained in Zehnder medium (Staub 1961) under the following conditions: temperature 22 ± 1 °C, illumination 20 mmol/m^2/s and light regime: 12 h light/12 h dark. The culture was treated against fungal contamination by washing three times with 100 mg/L cycloheximide (Sigma-Aldrich, Co., Saint Louis, MO, USA). Although the strain has been validly described as a new genus (Dvořák *et al.* 2014), we will treat it throughout the text as one of the *Synechococcus*-like lineages.

DNA was extracted from 50 mg of wet biomass using UltraClean Microbial DNA Isolation Kit (MOBIO, Carlsbad, CA, USA). A quality and concentration of DNA was evaluated via NanoDrop 1000 (Thermo Fisher Scientific, Wilmington, DE, USA) and separated on an ethidium bromide stained 1.5% agarose gel. The draft genome of strain *N. sphagnicola* CAUP A 1101 was obtained using pyrosequencing on a 454 GS Junior System (454 Life Sciences; a Roche company, Bradford, CT, USA). The shotgun sequencing library was prepared with following steps. Isolated DNA was nebulized in 30 psi for 1 min, and fragment ends were repaired using taq and T4 polymerase. Afterwards, the DNA was purified using Agencourt AMPure Beads XP system (Beckman Coulter, Beverly, MA, USA), RL adaptors were ligated, and small fragments were removed. Quality of the library was assessed on FlashGel system (Lonza, Basel, Switzerland). Gel was run for 6 min at 250 V. The library was quantified using TBS 380 Fluorometer (Topac, Cohasset, MA, USA) with PicoGreen dye (Topac). DNA fragments from the library were amplified on capture beads using emulsion PCR (polymerase chain reaction) using Live Amp Mix (454 Life Sciences, a Roche company). The sequencing run was performed for 200 cycles in the 454 GS Junior System platform.

A total of 331 932 reads in three runs with an average length of 361.1 bp were assembled de novo using the MIRA 4 assembler (Chevreux *et al.* 1999; parameters: -job=denovo,genome,accurate,454, -highlyrepetitive, -AS:nop=10). Contaminant contigs were identified using BLASTN against complete bacterial genomes with default parameters. All the contigs showing a match to a sequenced bacterial genome rather than cyanobacteria with a percentage identity higher than 90% and coverage higher than 90% were considered as contaminants and discarded (741 contigs). The assembled genome resulted in 118 contigs (>500 bp) with an N_{50} 81 579 bp, and a theoretical coverage of 22× based on the estimation of a length of 4.3 megabases calculated using MIRA assembler. Annotation was performed using The Rapid Annotation using Subsystems Technology (RAST) pipeline

(Aziz *et al.* 2008) with default options except enabled fix frameshift; tRNA was predicted using tRNAscan-SE 1.21 (Lowe & Eddy 1997; parameters: cove searching, covariance model—bacterial). Repeats were masked using REPEATMASKER 4.0.2 with default options (Smit *et al.* 2013). To compare repeats with other genomes, nine other cyanobacterial genomes (the closest neighbours in the RAST and in phylogenomic analysis) were selected and analysed with the same options. The genome has been deposited in GenBank (http://www.ncbi.nlm.nih.gov/).

16S rRNA phylogeny and molecular clocks

All available 16S rRNA sequences of cultured strains of *Synechococcus* (length > 900 bp) were downloaded from GenBank (http://www.ncbi.nlm.nih.gov/) by following query: '"*Synechococcus*"[Organism] OR Synechococcus [All Fields]) AND 16S[All Fields]) NOT uncultured[All Fields]'. Identical sequences were removed. One sequence of each *Synechococcus* clade was used for the most similar sequences identification using BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi). Only nonidentical sequences with minimal similarity 90% to the query were chosen. The last step involved addition of sequences of reference genomes and reference genera from the families Chroococcales, Oscillatoriales and heterocyst-forming genera (Nostocales). Total number of sequence at the end was 203 (Table S1, Supporting information).

Multiple sequence alignment was performed in MUSCLE 3.6 (Edgar 2004) with default options. All sequence positions within the alignment were analysed with *Gloeobacter violaceus* as the out-group. Phylogenetic reconstruction was performed using Bayesian maximum-likelihood inference implemented in MRBAYES 3.2.2 (Ronquist & Huelsenbeck 2003) via CIPRES science gateway (Miller *et al.* 2010). The most suitable substitution evolutionary model was identified using JMODELTEST 0.1.1 (Posada 2008) based on both Akaike and Bayesian criterion as GTR+G+I. Two parallel Markov chain Monte Carlo (MCMC) simulations were simultaneously run for 30 540 000 generations, each one with one cold and three heated chains. MCMCs were sampled every 5000th generation. Stop early if the convergence diagnostics falls below 0.01 was set to yes. The first 25% of trees were discarded as burn-in. The final consensus tree was constructed from all compatible groups, which produced a strictly bifurcating tree necessary for subsequent dating analysis.

The molecular chronogram was constructed using penalized likelihood (Sanderson 2002), implemented in r8s 1.8 (Sanderson 2012). Data analysis was performed on the consensus Bayesian tree constructed in MRBAYES (see detailed information above). The tree was calibrated with a combination of fossil and molecular calibrations previously published, mostly using calibrations by Shirrmeister *et al.* (2013). The root of the tree was calibrated using the fossil record. 3.8 BYA was used as an origin of existing life forms (Nisbet & Sleep 2001), and 2.7 BYA used as a date when oxygen-evolving cyanobacteria probably originated (Brocks *et al.* 1999). The origin of filamentous cyanobacteria was adopted from Shirrmeister *et al.* (2013); extreme values of their estimations were used (2.38–3.08 BYA). Third calibration point was the origin of akinetes, which was constrained with the fossil record (2.1 BYA; Amard & Bertrand-Sarfati 1997) and a molecular dating estimation (1.618 BYA; Falcon *et al.* 2010). Second analysis was performed in the same way, except minimum age for akinete node was set to 2.1 BYA (Amard & Bertrand-Sarfati 1997), which gives an interval of node ages. Rate variation was low in both analyses and did not affect branching order (for details see Table S2, Supporting information).

Phylogenomic analysis and molecular chronometer

Available and annotated genomes (both draft and complete) of *Synechococcus* were acquired from the ftp server of GENBANK (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/), database version 13 July 2013. Other genomes of Cyanobacteria from GENBANK were added to cover the broad evolutionary array of this group with a total number of 70, and representing most major niches/habitats. Genomes of *Leptolyngbya boryana* PCC 6306, *Geitlerinema* sp. PCC 7105 and *Leptolygbya* sp. PCC 7376 were re-annotated using RAST due to lack of annotation in the GenBank database (Table S3, Supporting information).

The super alignment for a subsequent phylogenomic reconstruction of a cyanobacterial species tree was obtained using phylogenomic Perl pipeline Hal (Robbertse *et al.* 2011) in the following steps: (i) orthologues were identified using all-vs-all BLASTP, a cut-off *e*-value of $1e^{-1}$, soft filtering of low-complexity regions was enabled, and segments of low complexity during the search phase were masked (Altschul *et al.* 1990). (ii) MCL clustering, cluster selection and filtering (van Dongen 2000). (iii) Multiple sequence alignment using MUSCLE 3.6 with default parameters, except an input order was set to stable (Edgar 2004). (iv) Alignment was edited using GBLOCKS (Castresana 2000) with half-liberal settings, and poorly aligned positions were discarded. (v) Concatenation of alignments of orthologous sequences into super alignment, see Robbertse *et al.* (2011) for details.

A resulting super alignment with a total of 46 978 amino acids was tested in PROTTEST 3.3 (Abascal *et al.*

2005) to find the most suitable substitution model based on both Akaike and Bayesian criterion (LG + G + I). The final phylogenetic tree was constructed in RAXML-HPC 8.0 (Stamatakis 2006) using the predicted model. *Gloeobacter violaceus* was used as an out-group. Tree topology was tested using rapid bootstrapping with 500 bootstrap replicates. This analysis was performed via CIPRES science gateway.

The final tree was also used for dating using molecular clocks. The same calibration points as in the 16S rRNA analysis were used. Rate variation was low in both analyses and did not affect branching order (for details see Table S2, Supplementary information).

Evolutionary events' analysis

Altogether, 192 orthologous genes' alignment was recovered from Hal. Divergent regions and poorly aligned positions were removed by GBLOCKS. An appropriate substitution model for each alignment was tested using PROTTEST based on both Akaike and Bayesian criterion. LG + I + G substitution model was the most suitable for 93.2% of alignments. Phylogenetic trees under maximum-likelihood criterion for each orthologue were estimated in PHYML 3.0 (Guindon & Gascuel 2003) using appropriate models. Reconciliation scenarios of HGT, gene loss, gene duplication and speciation were assessed and quantified using Python-based program ANGST (David & Alm 2010). Penalties were set as following: HGT (3.0), gene duplication (2.0), gene loss (1.0) and speciation (0.0). These penalty values seem to be the most suitable for prokaryotic organisms in general (David & Alm 2010). As a template, a species tree was constructed using maximum-likelihood RAXML based on super alignment (see estimation details above). The RAXML species tree was used as the reference tree for an ancestral state reconstruction of continuous data in MESQUITE 2.71 (Maddison & Maddison 2011), where the total number of HGT events were plotted over the species tree based on parsimony criterion.

Linear regression and correlation analyses were performed in PAST 3 (Hammer *et al.* 2001).

Results

Basic features of the new genome

The total length of the draft, near-complete genome was 4 331 368 bp with 50.14% G+C content. RAST genome annotation revealed a total of 4598 coding sequences (CDSs) and 48 RNAs. All common tRNAs were present as in other *Synechococcus* genomes. 53% of CDSs were annotated based on known proteins with biological function, and 47% were annotated as hypothetical proteins. The genome contained 182 simple repeats and 18 low-complexity repeats. All species contained mostly simple repeats and less low-complexity repeats (Fig. S1, Supporting information), except *Nostoc* sp. PCC 73102 which possessed one unclassified repeat according to Smit *et al.* (2013). More complex and larger genomes (heterocystous cyanobacteria) had more repeats in comparison with filamentous and unicellular.

Evolutionary history of Synechococcus

A 16S rRNA phylogeny based on all *Synechococcus* strains currently available in GENBANK (as of 19 September 2013) was constructed using Bayesian inference (Fig. 1). We noted at least 12 lineages, depending on the level of resolution corresponding to branch lengths, and *Synechococcus sensu lato* itself appears polyphyletic. Regardless of nodes recovered, lineages include a mixture of freshwater, marine, temperate and thermal isolates. While ecological specificity may be important for closely related taxa, it did not appear to be important in broad phylogenetic patterns for *Synechococcus*. Six lineages contained marine, five freshwater and three thermal isolates. Some lineages, such as clade 10 (Fig. 1), contained a mix of marine and freshwater picoplankton strains as well as some ecologically monophyletic clusters. Clade 12 is nested within the lineages that include the majority of other cyanobacterial lineages (Fig. 1, clade A). Several of the *Synechococcus* lineages were closely related or sister to other cyanobacteria, mostly *Leptolyngbya* with six clades, which were morphologically quite simple but were obligatorily filamentous. *Neosynechococcus sphagnicola* (its genome was sequenced in this work) was one of these 12 lineages; isolated from a peat bog (see Dvořák *et al.* 2014 for details), it clustered with the Antarctic strains of *Leptolyngbya frigida*. The other peat bog *Synechococcus* sp. PCC 7502 belonged to another clade 8 (Fig. 1).

A chronogram based on 16S rRNA sequence data revealed a great divergence of time between lineages (Fig. 1). The thermal strains appeared to be basal and radiated first to marine and then freshwater habitats, with subsequent reticulate ecological diversification (Fig. 1). Not all lineages were of equal evolutionary age; for example, clade 10, a mixture of freshwater and marine strains, appears to have evolved relatively recently compared to clade 1 (1.36–1.5 BYA). Only clades 1, 2 and 5 derived prior to the Great Oxidation Event (GOE, 2.32–2.45 BYA; Bekker *et al.* 2004). The most abundant *Synechococcus* marine picoplanktonic lineages (Fig 1, clade 10) have derived 1.81–2 BYA. This contrasted with trees constructed using whole genomes (Fig. 2), which provided an estimate of marine picoplanktonic *Synechococcus* of ca. 2.14–2.35 BYA. *N. sphagnicola* has derived 2.12–2.36 BYA.

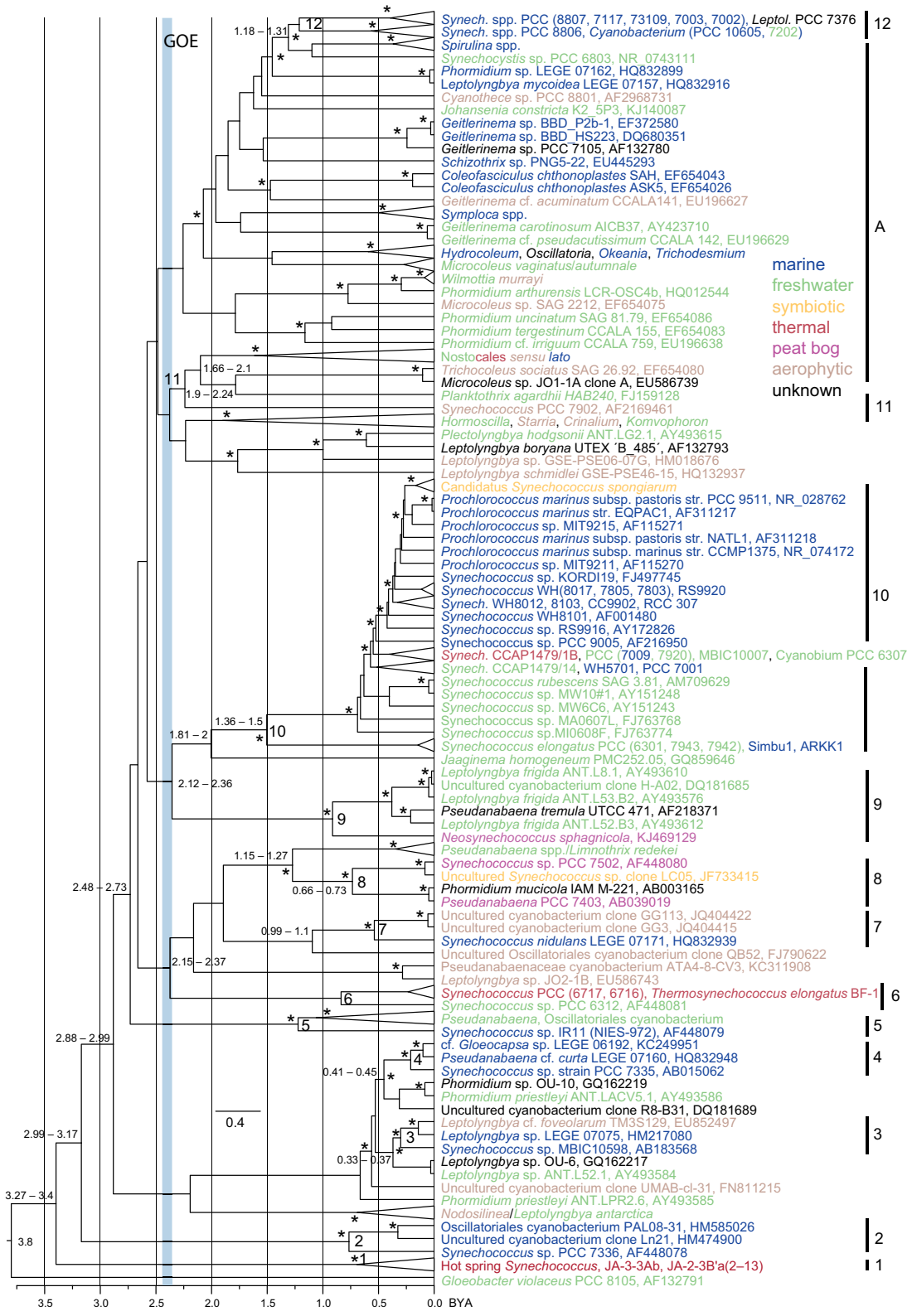

Fig. 1 A 16S rRNA chronogram of all *Synechococcus* lineages and other representative cyanobacteria based on Bayesian phylogeny. Both time estimates are combined in this figure. Asterisk represents posterior probabilities ≥0.9, and important ages are at the nodes with particular *Synechococcus* groups labelled. Habitats of strains are explained in the legend. Great Oxidation Event (GOE) is represented by the blue stripe. Isolation source of *Synechococcus*-like cyanobacteria is shown in the coloured legend.

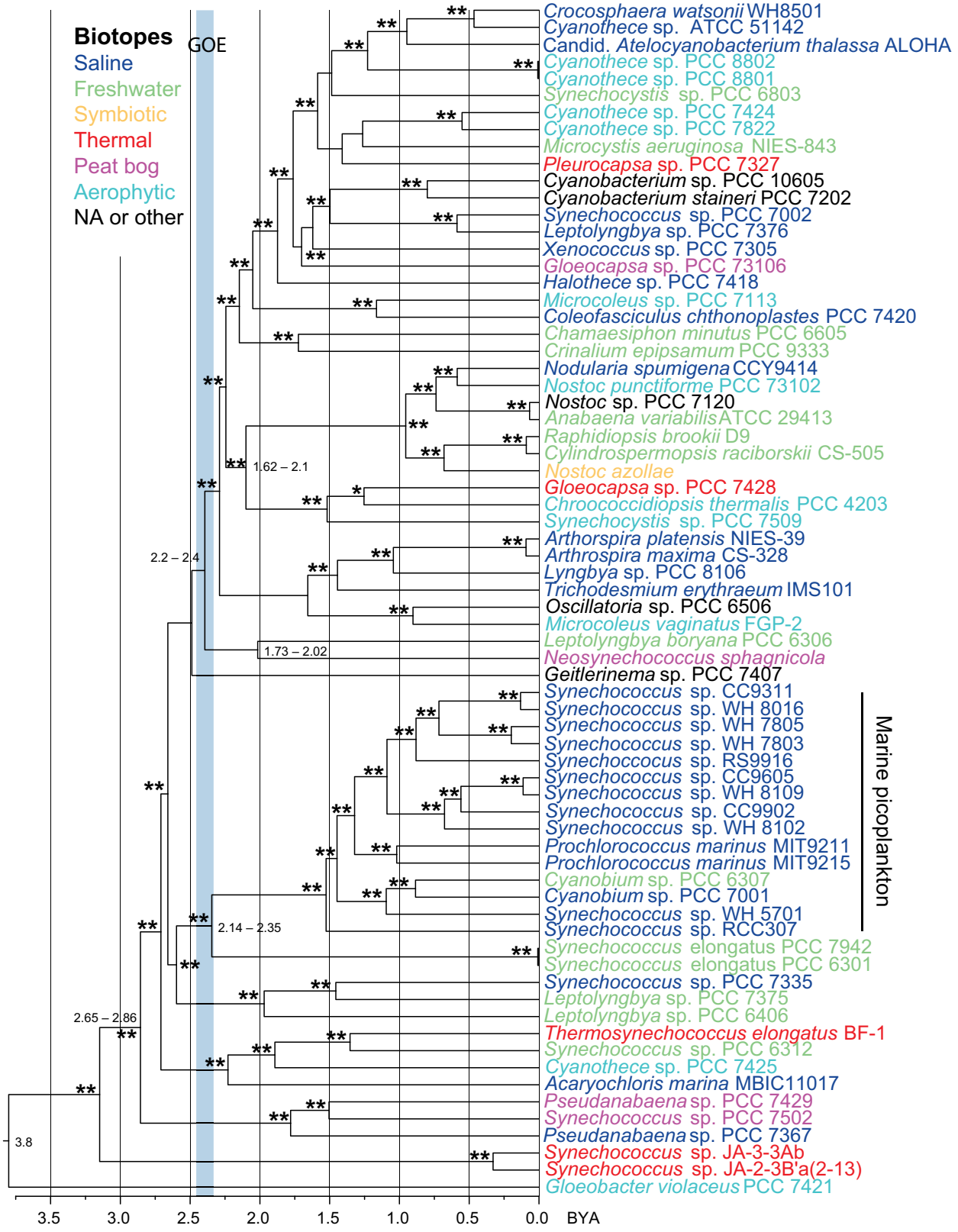

Fig. 2 A chronogram based on super alignment of 192 orthologous gene families of cyanobacterial genomes constructed using maximum-likelihood criterion. Both time estimates are combined in this figure. Bootstrap supports ≥70 are represented by asterisk and ≥90 by two asterisks together with the age at the nodes. Great Oxidation Event (GOE) is represented by the blue stripe.

However, trees constructed using either 16S rRNA or whole genomes recovered trees of very similar topology. For example, the thermal and marine picoplanktonic isolates and the derived clade 12 of the 16S rRNA tree are situated in the similar positions in genome phylogeny. There were six clades of *Synechococcus* recovered, because sequenced genomes offer only limited taxon sampling. Clades 3 and 4 (Fig. 1) were the putatively most recent radiations, having evolved ca. 0.48–0.5 BYA.

An analysis of HGT in Cyanobacteria

One hundred and ninety-two orthologous genes were analysed for HGT. Our analyses did not discern any patterns of HGT in relation to particular functional gene groups (see Table S4, Supporting information). For example, some ribosomal protein genes exhibited the lowest (17 events in 50S ribosomal protein L5) and at the same time highest (33 events in SSU ribosomal protein S17P) number of events.

Ancestral state reconstruction based on HGT events illustrates that the basal, thermophilic *Synechococcus* and *Pseudanabaena* lineages have the least number of HGT events (Fig. 3, clade 1, Table S5, Supporting information). Conversely, the most derived lineage has the greatest number of HGT events (Fig. 3, clade 2). This lineage contains taxa with no common morphological or ecological features, containing filamentous, unicellular, marine and freshwater strains. When a temporal scale was added (Fig. 3), it was apparent that the rate and tempo of HGT is highest at the onset of speciation among phylogenetically similar lineages, but tends to slow as diversification and evolutionary distance increases (Fig. 3). Estimated node ages and reconstructed number of HGT events on internal nodes were negatively correlated ($P < 0.001$, $r = -0.40835$). Speciation and HGT events of all investigated species were significantly correlated ($P < 0.001$, $r = 0.95132$). On the other hand, there was found no significant correlation between number of HGT events and genome size ($P = 0.34057$, $r = -0.11561$). Further, there appears to be no signal when morphological characters were mapped onto the tree (Fig. 3). Similar to the 16S rRNA tree, *Synechococcus* often fell out with *Leptolyngbya* and *Pseudanabaena*, all of which had roughly similar morphologies, differing mainly in filament formation. In total, *Synechococcus* lineages exhibited both the fewest and among the highest numbers of HGT, with individual lineages spanning the potential range of events. Our newly sequenced genome *N. sphagnicola* was sister to the genome of *Leptolyngbya boryana* PCC 6306 (clade 3), falling in the middle range of HGT compared to other *Synechococcus* lineages.

Discussion

Synechococcus belongs to one of the most important, yet enigmatic lineages of micro-organisms due to its worldwide distribution, primary productivity and ancient evolutionary origin (Honda *et al.* 1999; Robertson *et al.* 2001; Flombaum *et al.* 2013). We provide a first comprehensive study of all *Synechococcus*-like sequences using 16S rRNA phylogeny, phylogenomics and HGT analysis in all its lineages. Moreover, a new draft genome of *Synechococcus*-like cyanobacterium is presented, which helps to expand the evolutionary coverage of sequenced genomes and untangle evolutionary relationships within *Synechococcus*.

The draft genome of *Neosynechococcus* was undertaken as it possesses some unique life strategies, such as inhabiting hyaline cells of *Sphagnum*. This original isolate was very closely related to the filamentous cyanobacterium *Leptolyngbya* (see Dvořák *et al.* 2014 for details). Furthermore, based on morphology, ultrastructure, 16S rRNA, 16S-23S ITS and *rbc*L sequence, it has been recently described as a new, monospecific genus (Dvořák *et al.* 2014). Thus, it significantly expands knowledge of the enigmatic *Synechococcus* group on a genomic level. Moreover, this new isolate illustrates the urgent need to increase the evolutionary coverage within the cyanobacteria, because the majority of sequenced cyanobacteria are from culture collections, which represent only a small fraction of the biodiversity of cyanobacteria (Nabout *et al.* 2013).

The number of CDSs and genome size is positively correlated in bacteria, with ca. 1 coding sequence to 1 kb (Konstantinidis & Tiedje 2004; Lynch 2007). We note a similar relationship in *Neosynechococcus*. Similarly, tRNA transcripts were similar to other cyanobacterial genomes included in the genomic tRNA database (Chan & Lowe 2009). No elementary tRNAs were missing, and the composition and number of repeats was similar (Fig. S1, Supporting information). A genome annotation in bacteria based on homology reveals 30–70% of hypothetical proteins (Rost *et al.* 2003), which is also exhibited by the *Neosynechococcus* genome.

The 16S rRNA phylogeny is the most complete analysis of *Synechococcus* lineages to date. We show the polyphyletic nature of the marine lineages and recovered more lineages than previous authors (Honda *et al.* 1999; Robertson *et al.* 2001; see short overview in Table S6, Supplementary information). Moreover, we note repeated radiations from freshwater to marine and other habitats (e.g. the two peat-bog, thermal lineages). This is in contrast to other studies which noted more monophyletic lineages based on ecology, probably because the authors were working exclusively either with marine or freshwater strains (and the researchers

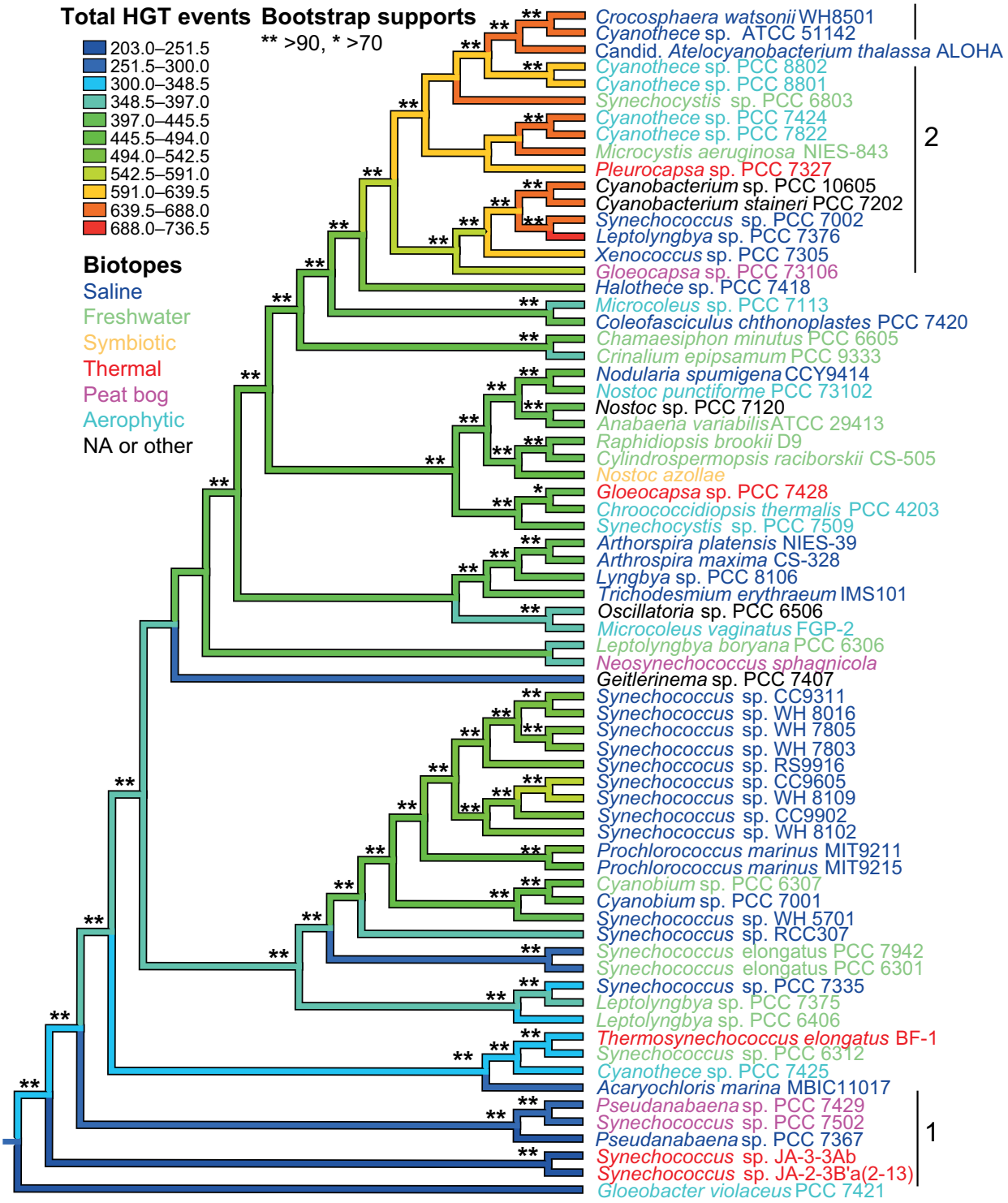

Fig. 3 The most parsimonious ancestral state reconstruction of a total number of HGT events for each taxon recovered from AnGST analysis. Bootstrap supports ≥70 are represented by asterisk and ≥90 by two asterisks. Habitat types and HGT levels are explained in the figure legend.

were often focusing on population diversity) or with limited taxon sampling (Rocap *et al.* 2002; Haverkamp *et al.* 2009; Callieri *et al.* 2013).

Based on our results, we do not consider *Synechococcus* as a monophyletic lineage. Further, there does not appear to be any morphological signal (in terms of cell

shape) in the lineages as evidenced by exemplars from each clade. We have confidence that the sequences from GenBank used in this study have been accurately described due to the simple morphology and relative ease of identification. Moreover, *Synechococcus*, *Pseudanabaena* and *Leptolyngbya*, although similar in 16S rRNA, exhibit very different ecology and morphology, and therefore, a classification based exclusively on 16S rRNA may lead to misidentification. This is particularly striking in the case of *Neosynechococcus sphagnicola*, where the closest sequence of filamentous Antarctic cyanobacterium *Leptolyngbya frigida* is 96.6% similar to *N. sphagnicola* 16S rRNA (Dvořák *et al.* 2014).

Cyanobacteria are a unique group among bacteria due to their high morphological diversity. However, there is actually a dearth of phylogenetically available morphological features to separate many cyanobacterial lineages, especially at the genus or species level (Dvořák *et al.* 2014). This lack of clear morphological signal presents problems when reconstructing ancestral states. For example, hypobradytely, the 'exceptionally low rate of evolutionary change exhibited by cyanobacterial taxa' (Schopf 1994), masks possible biochemical, ecological or genetic changes. The evolutionary history of cyanobacteria is a series of convergent evolutionary events leading to relatively sparse morphological features. Thus, the hypobradytely is only elusive, because fossil morphotypes are indistinguishable from recent morphotypes, but they have different evolutionary histories based on the phylogenomic and phylogenetic analysis.

Similar morphological and ecological (often indistinguishable) features have probably evolved multiple times on a great timescale. *Synechococcus* may be considered as an elementary state of the cyanobacterial morphological potential, where possibly every lineage may lead. However, there are significant differences in *Synechococcus* genomes (e.g. size, GC content, gene composition, see Table S3, Supporting information), which were sequenced quite recently (Dagan *et al.* 2012; Shih *et al.* 2013), indicating great changes in the genomes, but with similar resulting phenotype. Their complex evolutionary history is also showed by *Neosynechococcus*. Moreover, altogether, 12 lineages have derived in a polyphyletic manner and one lineage (Fig. 1, clade 12, Fig. 2) of *Synechococcus* has recently derived from the most complex lineages of filamentous cyanobacteria (Fig 1, clade A). Thus, we suggest that a possible splitting of the *Synechococcus* lineages into different genera is probable in the future.

Our analyses indicate that the main marine picoplanktonic lineages of *Synechococcus* evolved in the middle or after the GOE. This period of time was followed by the Paleoproterozoic 'Snowball Earth', a period of decreased temperature with subsequent increase to the recent values (Kopp *et al.* 2005). These conditions may have facilitated the rapid diversification of the marine lineages following possible mass extinctions which opened up new niche space. The marine *Synechococcus* (Fig. 1 clade 10, Fig. 2) has very long evolutionary history (up to 2.35 BYA) and, coupled with important roles in the global carbon cycle and oxygen production, ranks *Synechococcus* as among the most influential organisms in Earth's history. Flombaum *et al.* (2013) have posited that the importance of the marine *Synechococcus* will increase with global climate change, as increased temperatures will lead to a rise in abundance of marine picoplankton (*Synechococcus* and *Prochlorococcus*). It shows that temperature is an important factor shaping *Synechococcus* evolution on global scale; thus, it might have been one of the triggering factors for marine *Synechococcus* diversification.

However, the Snowball Earth environment was characterized by severe glaciation likely also in tropical oceans. Thus, where were *Synechococcus*-like cyanobacteria able to survive this period? It has been shown that thermal springs held microbial activity (Kirschvink *et al.* 2000). Thermal lineages of *Synechococcus* have diverged at least three times (Fig. 1), but their origin did not clearly coincide with the Snowball Earth events. During later Snowball Earth event in neo-Proterozoic, models suggest very thin ice shed or ice-free patches close to the equator, where cyanobacteria may have found refugia (Warren *et al.* 2002).

The majority of *Synechococcus* lineages (8 of the 12) diverged after the GOE, corresponding to a historic global temperature increase. It must be pointed out, however, that there exists a great degree of uncertainty in regard to the exact dating, mainly because cyanobacteria have a scant fossil record compared to plants and animals, but more than other prokaryotes (Schopf 2001). Moreover, the molecular dating of root has even higher degree of uncertainty (Sanderson 2012), but the earliest biomarker estimations place the origin of photosynthesis to 3.8 BYA (reviewed in Sleep 2010) while sulphur isotope analyses suggest a possible oxic atmosphere at 3.8 BYA (Ohmoto *et al.* 2006). Thus, it might be possible to place the origin of this clade to that time period, and *Synechococcus* before 3 BYA (Figs 1 and 2). Regardless, our analysis is largely in agreement with previously published estimations (Falcon *et al.* 2010; Shirrmeister *et al.* 2013). Thus, the radiation of a majority of cyanobacterial lineages diverged after the GOE. On the other hand, there are also later estimates of cyanobacterial origin (ca. 2.7 BYA) using more restricted roots when inferring molecular clocks (Hedges *et al.* 2001; Battistuzzi *et al.* 2004; Blank & Sanchez-Baracaldo 2010). For example, some authors suggest a freshwater origin

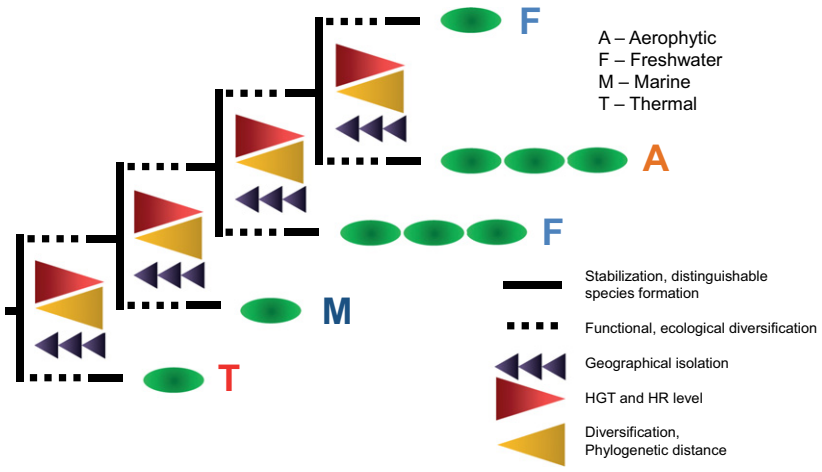

Fig. 4 A model of speciation in cyanobacteria constructed as an approximation of recent findings considering HGTs and HRs roles in speciation and data presented in this work. One green oval represents unicellular cyanobacterium and three ovals filamentous.

of cyanobacteria (Blank & Sanchez-Baracaldo 2010), which contradicts our 16S rRNA phylogeny (Fig. 1) and places the origin from thermal habitats with subsequent radiation to marine environments. However, Blank & Sanchez-Baracaldo (2010) only employed taxon with sequenced genomes and thus miss many potential taxa for their analyses. Moreover, they considered a basal *Synechococcus* isolate JA-3-3Ab and JA-2-3B'a(2-13) from the Octopus Spring in the Yellowstone Park, which has the same position in our phylogeny (Figs 1 and 2) as freshwater. It is questionable as to whether Octopus Spring can be considered as a freshwater ecosystem, because conductivity exceeds 3000 μS/cm (Havig 2009) and temperature 50 °C (Allewalt *et al.* 2006). Due to warmer climate of early Earth (reviewed in Kasting & Howard 2006), plausible origin of cyanobacteria can be placed in thermal habitats. Nevertheless, the most basal marine clade 2 in the 16S rRNA analysis (Fig. 1) does not have significant posterior probability at the node; thus, we cannot exclude any scenario at the moment.

Horizontal gene transfer (also LGT—lateral gene transfer) is considered one of the most important factors shaping prokaryotic species (Zhaxybayeva *et al.* 2006; Polz *et al.* 2013). Through HGT, novel genes might be acquired into the genome or homologous replacement might occur, which is revealed by analysis of reconciliation of gene trees compared to species trees (David & Alm 2010; Szollosi *et al.* 2012). HGT are present in both core genomes (genes dedicated primarily to basic metabolism and considered more stable; Shi & Falkowski 2008) and shell genomes (David & Alm 2010). The level of HGT significantly varies among cyanobacteria and among genes in the same genome (Fig. 3). Newly diverged lineages possess higher levels of HGT, while older lineages appear more stable. Of course, biases in taxon sampling might affect results due to oversampling of sequenced genomes in some environments (i.e. marine; Palenik *et al.* 2008). However, our analysis of HGT is the most exhaustive ever performed in cyanobacteria, considering the number of taxa analysed. On the other hand, there is no trend among genes analysed for HGT (e.g. ribosomal proteins). Any gene family with particular gene function does not seem to be preferred. Thus, it is probably that there is no selective pressure for some particular gene family exhibited by raised level of HGT.

Bacterial species concepts are subject to much debate and speculation (Johansen & Casamatta 2005; Achtman & Wagner 2008; Polz *et al.* 2013). Some researchers contend that there are no species in bacteria because high rates of gene loss and gain, coupled with HGT, would not lead to a cohesive evolutionary lineage, thus rendering species unrecognizable (Hanage *et al.* 2005). However, recent studies have shown that HGT and HR decrease with genetic distance and may be habitat specific (Popa *et al.* 2011; Smillie *et al.* 2011; Shapiro *et al.* 2012), although some infrequent HGT events are observed in distantly related taxa (Popa *et al.* 2011). Some authors propose that speciation in bacteria may be similar to 'sexual' taxa where the number of HR events is higher than mutation (Fraser *et al.* 2007). Therefore, barriers to gene flow exist, and thus, cohesive evolutionary units originate similar to plants or animals, but with different timescales and tempo and without actual sexual reproduction which is substituted by HGT and HR. Geography might also play an important role and give rise to allopatric speciation, which is not necessarily continuous and is on a larger geographical scale than animals and plants (Bahl *et al.* 2011; Mazard *et al.* 2011; Dvořák *et al.* 2012). Another explanation was offered by Polz *et al.* (2013), who suggested there exist local bacterial gene pools with high innovative genetic capabilities which may foster speciation rather than genetic isolation.

Speciation models show diversification in short time frames, typically acting on the population level (Shapiro

et al. 2012; Polz *et al.* 2013). We propose the model integrating speciation patterns already published (Shapiro *et al.* 2012; Polz *et al.* 2013) with the data in this study (Fig. 4); thus, we are able to present speciation events across long time periods and in many taxa, using easily traceable ecological and morphological phenotypic features. We used the genus *Synechococcus* because it is relatively simple to identify, although it is among the most polyphyletic lineages of cyanobacteria. Thus, it illustrates general trends in cyanobacterial (and possibly more broadly prokaryotic) evolution. For example, cyanobacteria seem to frequently switch between filamentous and unicellular forms, showing at least nine such switches in this study. Thus, they exhibit very frequent convergent evolutionary events. Moreover, clade 12 (Fig. 1) has derived from the most morphologically complex cyanobacteria. The majority of phenotypic traits recognized in cyanobacteria are not strictly shared or derived, but have been acquired (and lost) multiple times independently (cell morphology, thylakoid arrangement, colony formation, etc.), except for the formation of specialized cells (e.g. heterocysts and akinetes). However, some studies show possible multiple origins of the heterocystous cyanobacteria (Shirrmeister *et al.* 2013). Therefore, it is possible that most of the phenotypic characters are convergent, as similarity of convergent phenotypes is very high. However, this might be caused by the lack of resolution of techniques which are currently available. It has been suggested that a formation of pseudofilaments might be caused by one inactivated gene in *Synechococcus* (Miyagishima *et al.* 2005). Thus, these convergent events might be relatively frequent. A similar pattern in lifestyle of intracellular bacterial parasites was observed (Merhej *et al.* 2009). However, it should be pointed out that the filaments of *Leptolyngbya*, for example, is more complex than the pseudofilaments of *Synechococcus*, and thus, longer speciation would be expected.

Thus, we propose the following model of cyanobacterial speciation characterized by serial convergence of observed phenotypic features. When a new lineage diverges, there is high level of HGT and HR, which decrease with time and phylogenetic distance, until the species is stable (Popa *et al.* 2011; Polz *et al.* 2013). There might also occur periodic geographical isolation events on larger geographical scales (i.e. continents, Dvořák *et al.* 2012). At the initial stage, species might be considered as 'fuzzy' as they may be laden with conflicting phylogenies among gene families (Hanage *et al.* 2005). This stage does not necessarily lead to more complex or unique phenotypes (e.g. filament morphology, colony construction); the actual evolutionary pressures at this stage probably involve ecological conditions (Shapiro *et al.* 2012). After a time period, rates of HGT and HR decrease and the genome becomes more stable, resulting in less conflicting phylogenies (Polz *et al.* 2013). The species is then recognizable by molecular markers (Hanage *et al.* 2005). The resulting phenotype may be similar, more complex (e.g. filamentous) or less complex (e.g. unicellular). Then, the whole process may be repeated. Local bacterial gene pools *sensu* Polz *et al.* (2013) provide a support for our model, because they create pangenome of available genetic diversity, which is constantly changing through HGT, HR, selection and mutations, and it offers a place for acquisition of genes specific for a local environment. Considering high dispersal abilities of bacteria (Marshall & Chalmers 1997), all genes are probably transferred among distant location and they create the bacterial metapangenome. Local bacterial gene pools allow cyanobacteria with different evolutionary histories to have similar phenotypic features, because they evolved over a long duration with significant amounts of changes through local gene pools (Polz *et al.* 2013).

Synechococcus is clearly a successful lineage in terms of primary production, distribution and abundance. It has diverged 12 times over more than 3 billion years. We propose several factors that contribute to this global dominance. First, *Synechococcus* exhibits a rapid generation time (Moore *et al.* 1995). Second, the small size and shape of the cells allows *Synechococcus* to be competitive in terms of nutrient (Young 2007) and light acquisition (Morel *et al.* 1993). Third, the prevalence of HGT in *Synechococcus* lineages implies an excellent ability to receive and utilize exogenous DNA, putatively providing a selective advantage (this study and Palenik *et al.* 2008). In conclusion, *Synechococcus* is one of the most successful and influential organisms in Earth's history, with high impacts to the past, present and potentially future global environment.

Acknowledgements

This study was funded by following projects: ESF Post-UP II CZ.1.07/2.3.00/30.0041 and IGA UP Prf-2014001. We would like to thank to Walter Schuller (University of North Florida), who allowed us to use their computer server. Moreover, we would like to express our gratitude to the anonymous reviewers for their inspiring comments, which helped to improve the manuscript.

References

Abascal F, Zardoya R, Posada D (2005) PROTTEST: selection of best-fit models of protein evolution. *Bioinformatics*, **21**, 2104–2105.

Achtman M, Wagner M (2008) Microbial diversity and the genetic nature of microbial species. *Nature Review Microbiology*, **6**, 431–440.

Allewalt JP, Bateson MM, Revsbech NP, Slack K, Ward DM (2006) Effect of temperature and light on growth of and photosynthesis by *Synechococcus* isolates typical of those predominating in the Octopus spring microbial mat community of Yellowstone National Park. *Applied and Environmental Microbiology*, **72**, 544–550.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Amard B, Bertrand-Sarfati J (1997) Microfossils in 2000 ma old cherty stromatolites of the Franceville group, Gabon. *Precambrian Research*, **81**, 197–221.

Aziz RK, Bartels D, Best AA *et al.* (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*, **9**, 75.

Bahl J, Lau MCY, Smith GJD *et al.* (2011) Ancient origins determine global biogeography of hot and cold desert cyanobacteria. *Nature Communications*, **2**, 163.

Battistuzzi FU, Feijao A, Hedges SB (2004) A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evolutionary Biology*, **8**, 44.

Bekker A, Holland HD, Wang P-L *et al.* (2004) Dating the rise of atmospheric oxygen. *Nature*, **427**, 117–120.

Blank CE, Sanchez-Baracaldo P (2010) Timing of morphological and ecological innovations in the cyanobacteria – a key to understanding the rise in atmospheric oxygen. *Geobiology*, **8**, 1–23.

Brocks JJ, Logan GA, Buick R, Summons RE (1999) Archaean molecular fossils and the early rise of eukaryotes. *Science*, **285**, 1033–1036.

Callieri C, Coci M, Corno G *et al.* (2013) Phylogenetic diversity of nonmarine picocyanobacteria. *FEMS Microbiology Ecology*, **85**, 293–301.

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, **17**, 540–552.

Chan PP, Lowe TM (2009) GTRNADB: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Research*, **37**, D93–D97.

Chevreux B, Wetter T, Suhai S (1999) Genome sequence assembly using trace signals and additional sequence information. *Computer Science and Biology, Proceedings of the German Conference on Bioinformatics*, **99**, 45–56.

Dagan T, Roettger M, Stucken K *et al.* (2012) Genomes of stigonematalean cyanobacteria (subsection V) and the evolution of oxygenic photosynthesis form prokaryotes to plastids. *Genome Evolution and Biology*, **5**, 31–44.

David LA, Alm EJ (2010) Rapid evolutionary innovation during an Archaean genetic expansion. *Nature*, **469**, 93–96.

van Dongen S (2000) *Graph Clustering by Flow Simulation*. University of Utrecht, Utrecht.

Dvořák P, Hašler P, Poulíčková A (2012) Phylogeography of the *Microcoleus vaginatus* (cyanobacteria) from three continents – a spatial and temporal characterization. *PLoS ONE*, **7**, e40153.

Dvořák P, Hindák F, Hašler P, Hindáková A, Poulíčková A (2014) Morphological and molecular studies of *Neosynechococcus sphagnicola*, gen. et sp. nov. (Cyanobacteria, Synechococcales). *Phytotaxa*, **170**, 24–34.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797.

Erwin PM, Thacker W (2008) Cryptic diversity of the symbiotic cyanobacterium *Synechococcus spongiarum* among sponge hosts. *Molecular Ecology*, **17**, 2937–2947.

Falcon LI, Magallon S, Castillo A (2010) Dating the cyanobacteria ancestor of the chloroplast. *ISME Journal*, **4**, 777–783.

Flombaum P, Gallegos JL, Gordillo RA *et al.* (2013) Present and future global distributions of the marine cyanobacteria *Prochlorococcus* and *Synechococcus*. *Proceedings of the National Academy of Sciences*, **110**, 9824–9829.

Fraser C, Hanage WP, Spratt BG (2007) Recombination and the nature of bacterial speciation. *Science*, **315**, 476–480.

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, **52**, 696–704.

Hammer Ø, Harper DAT, Ryan PD (2001) PAST: paleontological statistics software package for education and data analysis. *Palaeontologia Electronica*, **4**, 9.

Hanage WP, Fraser C, Spratt BG (2005) Fuzzy species among recombinogenic bacteria. *BMC Biology*, **3**, 6.

Haverkamp THA, Schouten D, Doeleman M, Wollenzien U, Huisman J, Stal JL (2009) Colorful microdiversity of *Synechococcus* strains (picocyanobacteria) isolated for the Baltic Sea. *ISME Journal*, **3**, 397–408.

Havig JR (2009) Geochemistry of hydrothermal biofilms: composition of biofilms in a siliceous sinter-deposition hot spring. PhD Thesis, Arizona State University, Phoenix.

Hedges SB, Hsiong C, Kumar S, Wang DYC, Thompson AS, Watanabe H (2001) A genomic timescale for the origin of eukaryotes. *BMC Evolutionary Biology*, **1**, 4.

Holt JG, Krieg NR, Sneath PHA, Staley JT, Williamsn ST (1994) Group 11. Oxygenic phototrophic bacteria. In: *Bergey's Manual of Determinative Bacteriology*, 9th edn (ed. Holt JG), pp. 377–425. Williams & Wilkins, Baltimore.

Honda D, Yokota A, Sugiyama J (1999) Detection of seven major evolutionary lineages in cyanobacteria based on the 16S rRNA gene sequence analysis with new sequences of five marine *Synechococcus* strains. *Journal of Molecular Evolution*, **48**, 723–739.

Johansen JR, Casamatta DA (2005) Recognizing cyanobacterial diversity through adoption of a new species paradigm. *Algological Studies*, **117**, 71–93.

Kasting JF, Howard MT (2006) Atmospheric composition and climate on the early Earth. *Philosophical Transactions of the Royal Society B*, **361**, 1733–1742.

Kirschvink JL, Gaidos EJ, Bertani LE *et al.* (2000) Paleoproterozoic snowball Earth: extreme climatic and geochemical global change and its biological consequences. *Proceedings of the National Academy of Sciences*, **97**, 1400–1405.

Komárek J, Kopecký J, Cepák V (1999) Generic characters of the simplest cyanoprokaryotes *Cyanobium*, *Cyanobacterium* and *Synechococcus*. *Cryptogamie Algologie*, **20**, 209–222.

Konstantinidis KT, Tiedje JM (2004) Trends between gene content and genome size in prokaryotic species with larger genomes. *Proceedings of the National Academy of Sciences*, **101**, 3160–3165.

Kopp RE, Kirschvink JL, Hilburn IA, Nash CZ (2005) The Paleoproterozoic snowball Earth: a climate disaster triggered

by the evolution of oxygenic photosynthesis. *Proceedings of the National Academy of Sciences*, **102**, 11131–11136.

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, **25**, 955–964.

Lynch M (2007) Streamlining and simplification of microbial genome architecture. *Annual Review of Microbiology*, **60**, 327–349.

Maddison WP, Maddison DR (2011) *Mesquite: A Modular System for Evolutionary Analysis*. Version 2.75, http://mesquiteproject.org.

Marshall WA, Chalmers MO (1997) Airborne dispersal of Antarctic terrestrial algae and cyanobacteria. *Ecography*, **20**, 585–594.

Mazard S, Ostrowski M, Partensky F, Scalan DJ (2011) Multi-locus sequence analysis, taxonomic resolution and biogeography of marine *Synechococcus*. *Environmental Microbiology*, **14**, 372–386.

Merhej V, Royer-Carenzi M, Pontarotii P, Raoult D (2009) Massive comparative genomics analysis reveals convergent evolution of specialized bacteria. *Biology Direct*, **4**, 13.

Miller MA, Pfeiffer W, Schwartz T (2010) Creating the CIPRES science gateway for inference of large phylogenetic trees. In: *Proceedings of the Gateway Computing Environments Workshop* (eds Pierce M, Thomas M, Wilkins-Diehr N), pp. 1–8. IEEE, New Orleans.

Miyagishima S, Wolk CP, Osteryoung KW (2005) Identification of cyanobacterial cell division genes by comparative and mutational analyses. *Molecular Microbiology*, **56**, 126–143.

Moore LR, Goericke R, Chrisholm SW (1995) Comparative physiology of *Synechococcus* and *Prochlorococcus*: influence of light and temperature of growth, pigments, fluorescence and absorptive properties. *Marine Ecology Progress Series*, **116**, 259–275.

Morel A, Ahn Y-H, Partensky F, Vaulot D, Claustre H (1993) *Prochlorococcus* and *Synechococcus*: a comparative study of their optical properties in relation to their size and pigmentation. *Journal of Marine Research*, **51**, 617–649.

Nabout JC, da Silva Rocha B, Carneiro FM, Sant'Anna CL (2013) How many species of Cyanobacteria are there? Using a discovery curve to predict the species number. *Biodiversity and Conservation*, **22**, 2907–2918.

Nisbet EG, Sleep NH (2001) The habitat and nature of early life. *Nature*, **409**, 1083–1091.

Ohmoto H, Watanabe Y, Ikemi H, Poulson SR, Taylor BE (2006) Sulphur isotope evidence for an oxic Archean atmosphere. *Nature*, **442**, 908–911.

Palenik B, Ren Q, Tai V, Paulsen IT (2008) Coastal *Synechococcus* metagenome reveals major roles for horizontal gene transfer and plasmids in population diversity. *Environmental Microbiology*, **11**, 349–359.

Papke RT, Ramsin NB, Bateson MM, Ward DM (2003) Geographical isolation in hot spring cyanobacteria. *Environmental Microbiology*, **5**, 650–659.

Polz MF, Alm EJ, Hanage WP (2013) Horizontal gene transfer and the evolution of bacterial and archeal population structure. *Trends in Genetics*, **39**, 170–175.

Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T (2011) Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Research*, **21**, 599–609.

Posada D (2008) JMODELTEST: phylogenetic model averaging. *Molecular Biology and Evolution*, **25**, 1253–1256.

Robbertse B, Yoder RJ, Boyd A, Reeves J, Spatafora JW (2011) Hal: and automated pipeline for phylogenetic analyses of genomic data. *PLoS Currents*, **3**, RRN1213.

Robertson BR, Tezuka N, Watanabe M (2001) Phylogenetic analyses of *Synechococcus* strains (cyanobacteria) using sequences of 16S rDNA and part of the phycocyanin operon reveal multiple evolutionary lines and reflect phycobilin content. *International Journal of Systematic and Evolutionary Microbiology*, **51**, 861–871.

Rocap G, Distel DL, Waterbury JB, Chisholm SW (2002) Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. *Applied and Environmental Microbiology*, **68**, 1180–1192.

Ronquist F, Huelsenbeck JP (2003) MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

Rost B, Liu J, Nair R, Wrzeszcynski KO, Ofran Y (2003) Automatic prediction of protein function. *Cellular and Molecular Life Sciences*, **60**, 1637–2650.

Sanderson MJ (2002) Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Molecular Biology and Evolution*, **19**, 101–109.

Sanderson MJ (2012) *r8s version 1.8. Analysis of Rates ('r8s') of Evolution*. Section of Evolution and Ecology. University of California, Davis. http://loco.biosci.arizona.edu/r8s/.

Schopf JW (1994) Disparate rates, differing fates: tempo and mode of evolution changed from the precambrian to the phanerozoic. *Proceedings of the National Academy of Sciences*, **91**, 6735–6742.

Schopf JW (2001) The fossil record: tracing the roots of the cyanobacterial lineage. In: *The Ecology of Cyanobacteria: Their Diversity in Time and Space* (eds Whitton BA, Potts M), pp. 13–35. Springer, Berlin.

Shapiro BJ, Friedman J, Cordero OX *et al.* (2012) Population genomics of early events in the ecological differentiation of bacteria. *Science*, **336**, 48–51.

Shi T, Falkowski PG (2008) Genome evolution in cyanobacteria: the stable core and the variable shell. *Proceedings of the National Academy of Sciences*, **105**, 2510–2515.

Shih PM, Wu D, Latifi A *et al.* (2013) Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proceedings of the National Academy of Sciences*, **110**, 1053–1058.

Shirrmeister BE, Vos JM, Antonelli A, Bagheri HC (2013) Evolution of multicellularity coincided with increased diversification of cyanobacteria and the Great Oxidation Event. *Proceedings of the National Academy of Sciences*, **110**, 1791–1796.

Sleep NH (2010) The Hadean-Archean environment. *Cold Spring Harbor Perspectives in Biology*, **2**, a002527.

Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ (2011) Ecology drives a global network of gene exchange connecting the human microbiome. *Nature*, **480**, 241–244.

Smit AFA, Hubley R, Green P (2013) REPEATMASKER OPEN-4.0. http://www.repeatmasker.org.

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.

Staub R (1961) Research on physiology of nutrients of the planktonic cyanobacterium *Oscillatoria rubescens*. *Schweizerische Zeitschrift fur Hydrologie*, **23**, 83–198.

Szollosi GJ, Boussau B, Abby SS, Tannier E, Daubin V (2012) Phylogenetic modelling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proceedings of the National Academy of Sciences*, **43**, 17513–17518.

Usher KM, Fromont J, Sutton DC, Toze S (2004) The biogeography and phylogeny o unicellular cyanobacterial symbionts in sponges form Australia and the Mediterranean. *Microbial Ecology*, **48**, 167–177.

Warren SG, Brandt RE, Grenfell TC, McKay CP (2002) Snowball Earth: ice thickness on the tropical ocean. *Journal of Geophysical Research: Oceans*, **107**, 31-1–31-18.

Waterbury JB, Rippka R (1989) Subsection I. Order Chroococcales Wettstein 1924, emend. Rippka et al. (1979). In: *Bergey's Manual of Systematic Bacteriology* (eds Staley JT, Bryandt MP, Pfenning N, Holt JG), pp. 1728–1746. Williams & Wilkins, Baltimore.

Whitton BA, Potts M (2000) *The Ecology of Cyanobacteria. Their Diversity in Time and Space*. Springer, Berlin.

Young KD (2007) Bacterial morphology: why have different shapes? *Current Opinion in Microbiology*, **10**, 596–600.

Zhaxybayeva O, Gogarten JP, Charlebois RL, Doolittle F, Papke RT (2006) Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. *Genome Research*, **16**, 1099–1108.

P.D., V.O. and R.S. designed experiments. P.D. and R.S. analysed data. P.D., D.A.C, A.P., P.H. and R.S. wrote the manuscript.

Data accessibility

The genome assembly is available in GenBank under Accession no. JJML00000000. Additional information regarding the metadata included in this study is available in Tables S1, S2, S3, S4, S5 and S6 (Supporting information). Multiple sequence alignments and phylogenetic trees are stored: doi:10.5061/dryad.977k6.

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 A comparison of repeats present in *Neosynechococcus* genome with other cyanobacteria.

Table S1 A list of cyanobacterial 16S rRNA sequences used for phylogenetic inference.

Table S2 A rate variation recovered from r8s analysis.

Table S3 A list of analyzed genomes with their basic features.

Table S4 A list of evolutionary events of 192 orthologous genes reconstructed with ANGST.

Table S5 HGT and speciation events recovered using ANGST of each genome analyzed.

Table S6 A short overview of *Synechococcus* clades identified until now.